Automatic Extraction of References to Future Events from News Articles Using Semantic and Morphological Information

Yoko Nakajima†‡

Adviser Funito Masui†, Michal Ptaszynski† , Hiroshi Yamada†, Hirotoshi Honma‡

† Kitami Institute of Technology, ‡ National Institute of Technology, Kushiro College

Outline

I propose a method for **automatic extraction** of **future-reference sentences (FRS)**. In the method I apply both morphological and semantic information to represent sentences in **morphosemantic structure** and extract frequent patterns from FRS. Then, I perform a series of experiments, in which I firstly train fourteen classifier versions and compare them to choose the best one. I conclude that the proposed method is capable to automatically classify future-reference sentences, significantly outperforming state-of-the-art, and reaching **76% of F-score**.

Research Purpose

a probability of rain : 30%

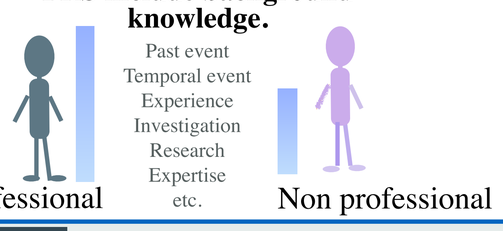
[bring / don't bring] an umbrella ?

professional : "It could rain around lunchtime a little."

laypeople : can decide to bring an umbrella easily.

- FRS on a specific topic can support decision making process.
- I undertake the task of extraction of FRS from reliable corpus (newspapers).

FRS include background knowledge.



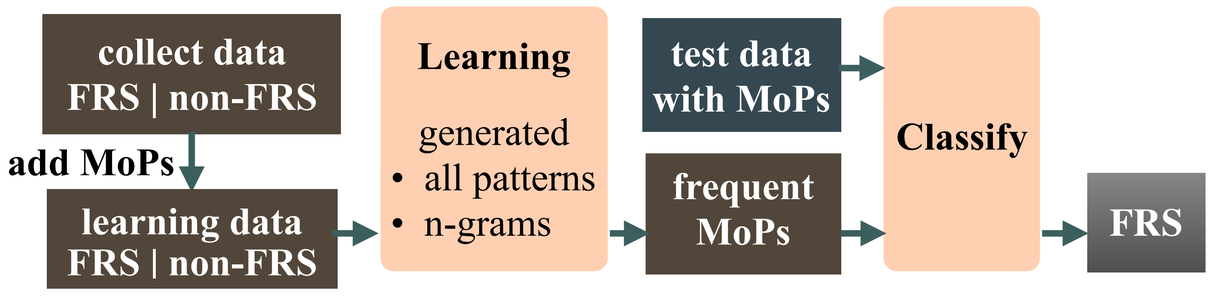
Previous Research

In my previous work I distinguished a variety of FRS.

- Future-reference expressions: 270 sentences from newspapers.
  time expressions: 70 / verbs: 141
- Probability of occurrence of future reference words in sentences
  one time: 45% , two times or more:55%

Classification of FRS with morphosemantic patterns (MoPs)

morphosemantic : morphology + semantic



Proposed Method

Morphosemantic Structure:

J: "*Nihon unagi ga zetsumetsu kigushu ni shitei sare, kanzen yoshoku ni yoru unagi no ryosan ni kitai ga takamatte iru.* "

E: "As Japanese eel has been specified as an endangered species, the expectations grow towards mass production of eel in full aquaculture. "

MoPs predicate argument structure
semantic role labelling : SRL

"[Object][Agent][State change][Action] [Noun] [State change][Object][State change] "

Argument Structure Analyzer [*1]

Thesaurus of predicate argument structure for Japanese verbs
words: 4400 semantic labels : 80 [*1Takeuchi et al. 2010]

Use SPEC for training and classification:
Sentence Pattern Extraction arChitecture [*2]

- **Sophisticated patterns** (with disjoint elements)
  - awarding length (LA)
  - awarding length and occurrence (LOA)
  - awarding none (normalized weight, NW)
  - using all patterns (ALL)
  - erasing all ambiguous patterns (AMB)
  - erasing only those ambiguous patterns which appear in the same number in both sides (zero patterns, 0P)
  - patterns (PAT)
  - only n-grams (NGR)
- **n-fold cross validation**
- **Results calculated in F-score, Precision, Recall**
- **Choose the most useful pattern**

[*2 Ptaszynski et al. 2011]

Experiment Setup

Extracting frequent MoPs

- Japan economy / Asahi(national) / Hokkaido(local) newspapers
- 1000 sentences extracted randomly
- annotate FRS or other
- by one expert and two lay people

training data
set50 and **set130**
FRS (50 or130) and other sentences (50 or 130)

Classification

Experiment1: verification of **frequent MoPs**
Experiment2: validation of **fully optimized model**

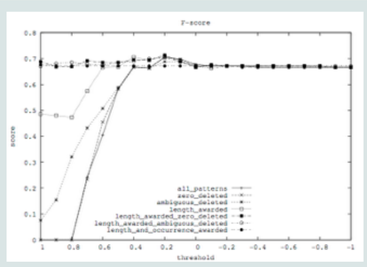
- Mainichi Newspaper (1996)
  topics: economy, international event, energy
  270 sentences
- by one expert and two lay people

Results of experiments

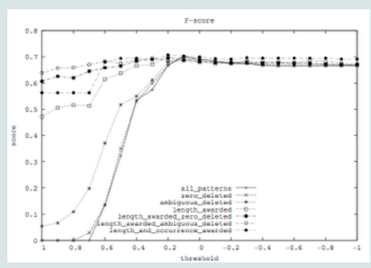
Extracting frequent MoPs

- learning data: set50, set130
- 10-fold cross validation

Compare to F-scores set130 and set50



F-score with set50



F-score with set130

Classification

Experiment1:

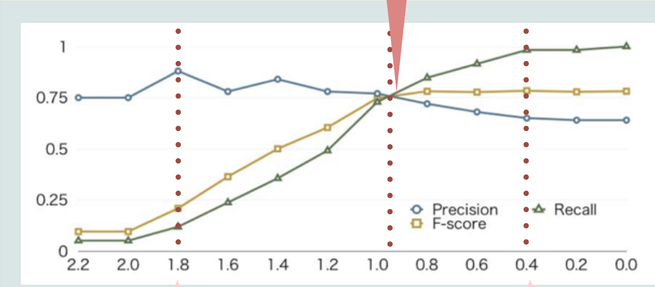
Accuracy of extracted FRE with frequent patterns

Pattern set	P	R	F
A: 10 patterns	0.39	0.49	**0.43**
B: 15 patterns	0.38	0.49	**0.43**
C: 5 patterns	0.35	0.35	0.35
D: 10 patterns with only over 3 elements	**0.42**	0.37	0.40
10 temporal expressions [*3]	0.50	0.05	0.10

[*3 Jatowt and Au Yeung 2011]

Experiment2:

break-even point
0.76 at 0.98



P=0.89
R=0.13
F=0.22

P=0.65
R=0.98
F=0.78

Extracted FRS

1. score=2.27

RJ: ***Dosha*** *wa kore made,****Shigen Enerugi Cho ni taishi****, , do hatsudensho no* ***heisa****, kaitai ni tsuite hoshin o setsumei shitekitaga, kaitai ni tsuite no hoteki kisei wanai tame, dochō mo kaitai no kettei o shitatameru koto* ***ni nari so da****.*

E : So far the **company** has been describing to **the Agency for Natural Resources and Energy** the policy for either **closure** or dismantling of the plant, and since there are no legal regulations found for dismantling, it is most likely that the agency **will also lean to** the decision of dismantling.

MS: **[Agent]** [Other] **[Organization]** [Action] **[State-change]** [State-change][Object][Role] [State-change] [State-change][Action][Adjective][Thing][Agent][State-change][Other] **[Verb]**

MoPs : [Agent]*[Verb],
[Agent]*[Organization]*[Verb],
[Agent]*[Action][State-change]*[Verb],
[Agent]*[Organization]*[State-change]*[Verb] .

I am glad to hear your suggestions and comments.

Conclusion

- We extracted FRS with morphosemantic .
- We verified validity of FRS patterns by two experiments.
  1. using only 5-15 frequent patterns (F-score = 0.43)
  2. validated classification for fully optimized model
     (F-score = 0.78, P = 0.65, R = 0.98)

Future work

- apply to real world events on large data.
- validate supporting to decide "yes" or "no" a future event with FRS.
  - ranking FRS
  - ordering timeline (short span, long span)